# 北 京 邮 电 大 学

# 本科毕业设计（论文）开题报告

| | | | | | |
|---|---|---|---|---|---|
| 学院 | 计算机学院（国家示范性软件学院） | 专业 | 智能科学与技术 | 班级 | 2019211315 |
| 学生姓名 | 张梓靖 | 学号 | 2019211379 | 班内序号 | 27 |
| 指导教师姓名 | 王纯 | 所在单位 | 计算机学院（国家示范性软件学院） | 职称 | 高级工程师 |
| 设计(论文)题目 | （中文）一种基于工作量的 Serverless 计算自动伸缩算法的设计与实现 | | | | |
| | （英文）Design and Implementation of Workload-based Auto-scaling Algorithm for Serverless Computing | | | | |

## 一、 选题背景及意义

Serverless 计算是近年来比较流行的一种云计算模型，它提供了一种无服务器的计算方式，可以大大降低企业运维成本，提高系统的弹性和可用性。但是，随着业务的发展，Serverless 服务的负载情况也会发生变化，有时会出现突发流量的情况，导致系统资源不足，影响服务的正常运行。为了解决这一问题，我们计划开发一种基于工作量的 Serverless 计算自动伸缩算法，使用前沿机器学习方法如 LSTNet、TPA-LSTM 等，根据 Serverless 服务的历史负载情况，进行时间序列预测，从而实现更加精准、节约、高效的自动伸缩。

Serverless 是云计算的一种设计思想，它的特点在于，不需要用户持续维护服务器、操作系统以及代码运行所需的基本环境，而是将这些前置需求部署到云端。这样，用户可以更专注于开发和部署应用，而不必担心底层基础架构的管理和维护。

Serverless 计算的主要特点包括：

- 按需执行：代码会在请求到来时被触发执行，而不是持续运行，从而有效降低计算成本。
- 无服务器：用户不需要维护服务器，也不需要关心服务器的配置和管理。
- 可扩展性：在需要时自动增加或减少计算资源，以应对流量高峰期或突发事件。
- 简单开发：在 Serverless 计算中，用户只需要关注应用的业务逻辑，而无需

关心底层架构。这样可以大大简化开发过程，并且能够更快地完成应用的部署和扩展。

## 二、 研究的基本内容

虽然自动伸缩技术在 Serverless 计算中发挥了重要作用，但它目前也存在一些问题。

一方面是自动伸缩算法的精确度。目前自动伸缩算法大多是基于某些预定义的性能指标或阈值来决定扩展或缩减计算资源，但这些指标和阈值并不能完全反映应用的实际需求。因此，一些自动伸缩算法可能会导致过度扩展或过度缩减计算资源，从而影响应用的性能和可用性。

另一方面，自动伸缩技术的可靠性也是一个问题。目前自动伸缩技术的可靠性并不完美。折衷不完美是两方面的，一种是伸缩过程的不完美，比如可能会导致系统故障或自动伸缩失败，从而影响应用的性能和可用性。另一种是伸缩策略的不完美，例如依赖于历史周期的伸缩，能否良好适应突发情况，是需要考虑的。

本项目主要针对前者，即算法的精确度尝试进行改进。此外，目前基于机器学习或传统学习、预测方法的自动伸缩算法往往能将周期性把握得很好，但对于长期趋势的把握可能不够。实际上随着业务的发展，有可能在存在日、周、月度周期的同时，还存在整体的业务上升趋势，从而急切需要一个更有适应性的模型。

## 三、 研究方法及措施

我们试图探索一种方法，能够不但捕捉到容器伸缩的时间周期，也能捕捉到伸缩随着业务发展的长期趋势。

在更广泛的意义上，这一项目也是在推动机器学习技术在自动伸缩领域的应用，为相关行业提供更多的选择和便利。同时，通过对 Serverless 计算负载情况的深入研究和分析，我们还可以为相关行业提供更为专业的建议和解决方案。

同时，本项目还会研究如何推进具体技术的落地。因为目前机器学习算法在落地过程中，往往发现有推理速度慢、资源消耗高、运行条件苛刻的情况。所以我们也会研究如何通

过一些先进的技术，例如深度学习编译，模型优化等技术，来推进算法的落地。

对于一个发展期的业务，不但存在短期的周期性（例如日、周访问量会有明显的周期），往往还存在长期的上升趋势（例如这个月 DAU 等指标，比上个月要高出一定数值，或是一定比例），所以长期上有可能表现出线性或者指数型增长。

在这种情况下，由于时间序列数据具有周期性和长期上升趋势，因此可以使用深度学习模型或者结合深度学习模型和线性模型的方法来进行预测。例如 LSTNet/TPA-LSTM/TCN 模型。

对于一个稳定期的业务，可能只存在短期的周期性，而长期，例如月、年的尺度上，由于主要是存量市场，用户增长和流失的速率基本平衡，运营也相对稳定，导致流量变化不显著。由于时间序列数据只具有短期的周期性，长期尺度上流量变化不显著，因此可以使用基于时间特征的线性回归模型或者 ARIMA 模型来进行预测。

**四、 研究工作的步骤与进度**

2023.1.1 ~ 2023.2.10 完成领域内容调研，论文对应部分撰写。
2023.2.28~2023.4.15 完成相关研究，设计程序。
2023.4.16~2023.4.30 进行设计评估和比较分析。
2023.5.1~2023.5.15 论文整体撰写。

**五、 主要参考文献**

[1] Kubernetes Authoritative guide version 4, author: Zheng Gong, Zhihui Wu, Xiulong Cui, Jianyong Yan.

[2] Docker technology introduction, author: Baohua Yang.

[3] Kubernetes docs: https://kubernetes.io/docs/home/

[4] OpenFaaS docs: https://docs.openfaas.com/

[5] Eric Jonas, Johann Schleier-Smith, Vikram Sreekanti and et al., “Cloud programming simplified: A berkeley view on serverless computing,“ arXiv preprint arXiv:1902.03383, 2019.

[6] Laszlo Toka, Gergely Dobreff, Balazs Fodor and Balazs Sonkoly, “Adaptive AI-based auto-scaling for Kubernetes,“ in 2020 20th IEEE/ACM International Symposium on Cluster, Cloud and Internet

Computing (CCGRID). IEEE, 2020, pp. 559-608.

[7] BINGO Hong, 时间序列预测方法总结: https://zhuanlan.zhihu.com/p/67832773

| | | | |
|---|---|---|---|
| 指导教师签字 |  | 日期 | 2023 年 1 月 1 日 |

注：可根据开题报告的长度加页。